DOCUMENT RESUME

ED 115 662 TM 004 932

AUTHOR Donlon, Thomas F.
TITLE An Optimizing Weight For Wrong Scores.
PUB DATE [May 75]
NOTE 19p.; Paper presented at the Annual Meeting of the New England Educational Research Organization (7th, Provincetown, Massachusetts, May 1-3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS Achievement Tests; Comparative Analysis; Guessing (Tests); *Multiple Choice Tests; *Response Style (Tests); *Scoring Formulas; Statistical Analysis; Testing; *Test Reliability; Test Validity; True Scores; *Weighted Scores

ABSTRACT

This study empirically determined the optimizing weight to be applied to the Wrongs Total Score in scoring rubrics of the general form = R - kW, where S is the Score, R the Rights Total, k the weight and W the Wrongs Total, if reliability is to be maximized. As is well known, the traditional formula score rests on a theoretical framework which is of dubious validity. Two instruments, variant approaches to the assessment of mathematical knowledge, were administered to approximately 1,700 entering college freshmen during an orientation period. The method consists of an iterative computer procedure for calculating split-half reliability of the tests as the weights are systematically varied throughout the region of maximization as determined by essentially canonical approaches. The results indicate that in contrast to the negative weight for the a priori formula score, a sizable positive weight maximizes reliability. The implications for rate of work as the single most reliable aspect of test performance seem clear. The validity of much educational testing rests on assumptions of fairness to those tested, achieved through optimization of standardized conditions. The study suggests that factors which alter rate-of-work characteristics of performance may be most detrimental to candidate success. (Author/DEP)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED115662

An Optimizing Weight for "Wrong" Scores

Thomas F. Donlon
Educational Testing Service
Princeton, New Jersey

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Seventh Conference of New England
Educational Research Organization
Provincetown
May 1-3, 1975

TM004 932

An Optimizing Weight for "Wrongs" Scores*

In scoring a multiple-choice test, the "formula score" or "correction for guessing" is the most widely used alternative to the simple count of the total number of right answers. The formula is

$$F.S. = R - \frac{W}{k-1},$$

where

F.S. = Formula Score
R = Total number right
W = Total number wrong
k = Number of choices per test item

The basic assumption which underlies this formula is that responses fall into two categories: those based on knowledge sufficient to determine a correct answer, and those based on knowledge insufficient to provide any basis for response better than chance responding. The value of R, the total number right, is a combination of the two categories, but the value of W, the total number wrong, reflects only responses based on insufficient information. The size of W is used to "correct" the observed value of R, to estimate the true value of the number of responses based on knowledge, for the chance behaviors are assumed to be randomly spread equally across the k choices per item, so that $\frac{k-1}{k}$ of them will be wrong answers, summing to the observed W score, and $\frac{1}{k}$ of them will be right answers, "buried" in the R score. The ratio of "buried" wrong answers to "observed" wrong answers is thus $\frac{1}{k-1}$.

Thorndike (1971) has discussed this correction, emphasizing its logical flaws and some of its merits. Ebel (1972) has presented research

*The author is indebted to David R. Saxe for the computer operations and for his interest in the problem.

evidence on the superior reliability of tests when they are scored with a formula correction. More recently, Lord (1975) has focussed on examinee behavior under different sets of instructions: formula scoring directions and number-right directions. He states an assumption that under number-right scoring candidates replace "Omit" responses by random marks on the answer sheet. The impact of this random responding is to reduce the sampling error of the formula score when contrasted with the number right score. This point is established by considering not $F.S. = R - \left(\frac{1}{k-1}\right) W$, but $F.S.' = R + \frac{O}{k}$, where O = the number of items unanswered, and R and k are as before. It has long been known that since R, W, and O sum to a constant, (T, the total number of items) the two values of the formula scores, F.S. and F.S.', are perfectly correlated.

But the assumption of random responses is not an attractive one. Lord is clearly concerned that the assumption be recognized for its crucial role and that instructions be developed to insure that any omissions under formula scoring are truly items for which candidates have only a chance, random, potential for success. But the theory is not strongly substantiated by our evidence on candidate behavior. Guessing on tests is in the main not random activity.

If the theoretical underpinnings of the formula score are so unattractive, why are we constrained to the weight, $\frac{1}{k-1}$, which it leads to for W? What other weight might we use, and to what purpose? One purpose might clearly be the development of a maximum reliability for the score from a test. In an unpublished study by Fischer and Jackson (1971),

the maximization of reliability was taken as the rationale for determining the best weight, x, for the wrongs. Taking Dressel's (1940) formula for the Kuder-Richardson reliability of a formula-scored test, Fischer and Jackson differentiated the equation with respect to the weights for the wrong answers when the right answers are weighted unity. That is, defining a weighted score as

$$W.S. = R + xw$$

where x may take any value, positive or negative, for what value is the reliability of the W.S., the weighted score, a maximum?

Somewhat to their surprise the authors found that the value of x was positive; the sum of the rights and a fraction of the wrongs was the most reliable score. Further, the Rights score alone was more reliable than the conventional formula score in each of four separately--timed subtests, comprising a form of the College Board Scholastic Aptitude Test (SAT), were two verbal and two mathematical sections with x-values of $+ .295$ and $+ .585$ for the mathematical material and $+ .639$ and $+ .720$ for the verbal.

Lord, in discussing this result observed that "This does not mean that we should give bonuses for wrong answers. It merely means that that trait of omitting items is a trait that can be quite reliably measured." This trait of omitting items, however, may be the trait of working on test material with a consistent speed. Lord, states in his discussion that his theoretical development will work best for unspeeded tests. But the test studied by Fischer and Jackson was a standard SAT form, moderately speeded. There is a possible difference between omitting an item and

not reaching it. In the standard ETS item analysis, an item is considered omitted if there is a response to a later item; it is considered Not Reached if there are no responses to later items. If the preponderance of omitting in Fischer and Jackson's paper was due to a failure to complete the test, to Not Reaching, this would be evidence that the trait which is reliably measured is rate of work, not tendency to omit due to conservatism or caution.

Fischer and Jackson used a generalized internal consistency approach, via Dressel's formula, and determined the maximum reliability by differentiation with respect to the weight for wrongs. The present study extends this work by an empirical determination of the correlation between two half tests on two 50-item mathematics tests. Each half test was scored $R + kW$, (k here is simply the weight in wrongs, exactly equivalent to Fischer and Jackson's x) and the correlation between them computed. This was systematically followed throughout the region $-5.0 < k < 5.0$. The result was the two empirical curves presented in Figure 1 and Figure 2. Each of these curves shows a maximum for a positive weight somewhat less than unity. Tables 1 and 2 provide the data upon which the graphs were based.

This result supports the finding of Fischer and Jackson. The two curves reflect slightly different treatments, however. The curve in Figure 1 was based on a 50-item mathematical test which consisted of data sufficiency items. The curve in Figure 2 is based on a 50-item mathematical test which consisted of "regular math" problems. The data sufficiency items have the form of two statements and a question. The respondent is

BEST COPY AVAILABLE

Table 1

Interform Reliability (R) of the Score R + kW
for Selected Values of k : Data Sufficiency Tests

(k)	(R)		(k)	(R)		(k)	(R)	
-0.500000D 01	0.611740D 00	1	-0.160000D 01	0.615620D 00	35	0.190000D 01	0.702700D 00	69
-0.490000D 01	0.611720D 00	2	-0.150000D 01	0.616260D 00	36	0.190000D 01	0.693670D 00	70
-0.480000D 01	0.611710D 00	3	-0.140000D 01	0.617020D 00	37	0.200000D 01	0.685920D 00	71
-0.470000D 01	0.611700D 00	4	-0.130000D 01	0.617900D 00	38	0.210000D 01	0.679250D 00	72
-0.460000D 01	0.611690D 00	5	-0.120000D 01	0.618940D 00	39	0.220000D 01	0.673490D 00	73
-0.450000D 01	0.611690D 00	6	-0.110000D 01	0.620160D 00	40	0.230000D 01	0.668490D 00	74
-0.440000D 01	0.611680D 00	7	-0.100000D 01	0.621600D 00	41	0.240000D 01	0.664140D 00	75
-0.430000D 01	0.611680D 00	8	-0.900000D 00	0.623320D 00	42	0.250000D 01	0.660330D 00	76
-0.420000D 01	0.611680D 00	9	-0.800000D 00	0.625360D 00	43	0.260000D 01	0.656980D 00	77
-0.410000D 01	0.611690D 00	10	-0.700000D 00	0.627820D 00	44	0.270000D 01	0.654020D 00	78
-0.400000D 01	0.611690D 00	11	-0.600000D 00	0.630780D 00	45	0.280000D 01	0.651400D 00	79
-0.390000D 01	0.611700D 00	12	-0.500000D 00	0.634370D 00	46	0.290000D 01	0.649050D 00	80
-0.380000D 01	0.611720D 00	13	-0.400000D 00	0.638750D 00	47	0.300000D 01	0.646960D 00	81
-0.370000D 01	0.611740D 00	14	-0.300000D 00	0.644110D 00	48	0.310000D 01	0.645070D 00	82
-0.360000D 01	0.611760D 00	15	-0.200000D 00	0.650700D 00	49	0.320000D 01	0.643370D 00	83
-0.350000D 01	0.611790D 00	16	-0.100000D 00	0.658830D 00	50	0.330000D 01	0.641820D 00	84
-0.340000D 01	0.611820D 00	17	0.0	0.669860D 00	51	0.340000D 01	0.640420D 00	85
-0.330000D 01	0.611870D 00	18	0.100000D 00	0.681210D 00	52	0.350000D 01	0.639140D 00	86
-0.320000D 01	0.611920D 00	19	0.200000D 00	0.696280D 00	53	0.360000D 01	0.637970D 00	87
-0.310000D 01	0.611980D 00	20	0.300000D 00	0.714310D 00	54	0.370000D 01	0.636900D 00	88
-0.300000D 01	0.612050D 00	21	0.400000D 00	0.735210D 00	55	0.380000D 01	0.635910D 00	89
-0.290000D 01	0.612140D 00	22	0.500000D 00	0.758140D 00	56	0.390000D 01	0.635000D 00	90
-0.280000D 01	0.612230D 00	23	0.600000D 00	0.781190D 00	57	0.400000D 01	0.634160D 00	91
-0.270000D 01	0.612350D 00	24	0.700000D 00	0.801350D 00	58	0.410000D 01	0.633380D 00	92
-0.260000D 01	0.612480D 00	25	0.800000D 00	0.815150D 00	59	0.420000D 01	0.632660D 00	93
-0.250000D 01	0.612630D 00	26	0.900000D 00	0.820160D 00	60	0.430000D 01	0.631980D 00	94
-0.240000D 01	0.612800D 00	27	0.100000D 01	0.816080D 00	61	0.440000D 01	0.631360D 00	95
-0.230000D 01	0.613010D 00	28	0.110000D 01	0.804600D 00	62	0.450000D 01	0.630770D 00	96
-0.220000D 01	0.613240D 00	29	0.120000D 01	0.789210D 00	63	0.460000D 01	0.630220D 00	97
-0.210000D 01	0.613510D 00	30	0.130000D 01	0.771990D 00	64	0.470000D 01	0.629700D 00	98
-0.200000D 01	0.613810D 00	31	0.140000D 01	0.755030D 00	65	0.480000D 01	0.629220D 00	99
-0.190000D 01	0.614170D 00	32	0.150000D 01	0.739350D 00	66	0.490000D 01	0.628760D 00	100
-0.180000D 01	0.614580D 00	33	0.160000D 01	0.725390D 00	67	0.500000D 01	0.628330D 00	101
-0.170000D 01	0.615060D 00	34	0.170000D 01	0.713210D 00	68			

BEST COPY AVAILABLE

Table 2

Interform Reliability (R) of the Score R + kW
for Selected Values of k : Math Achievement Test

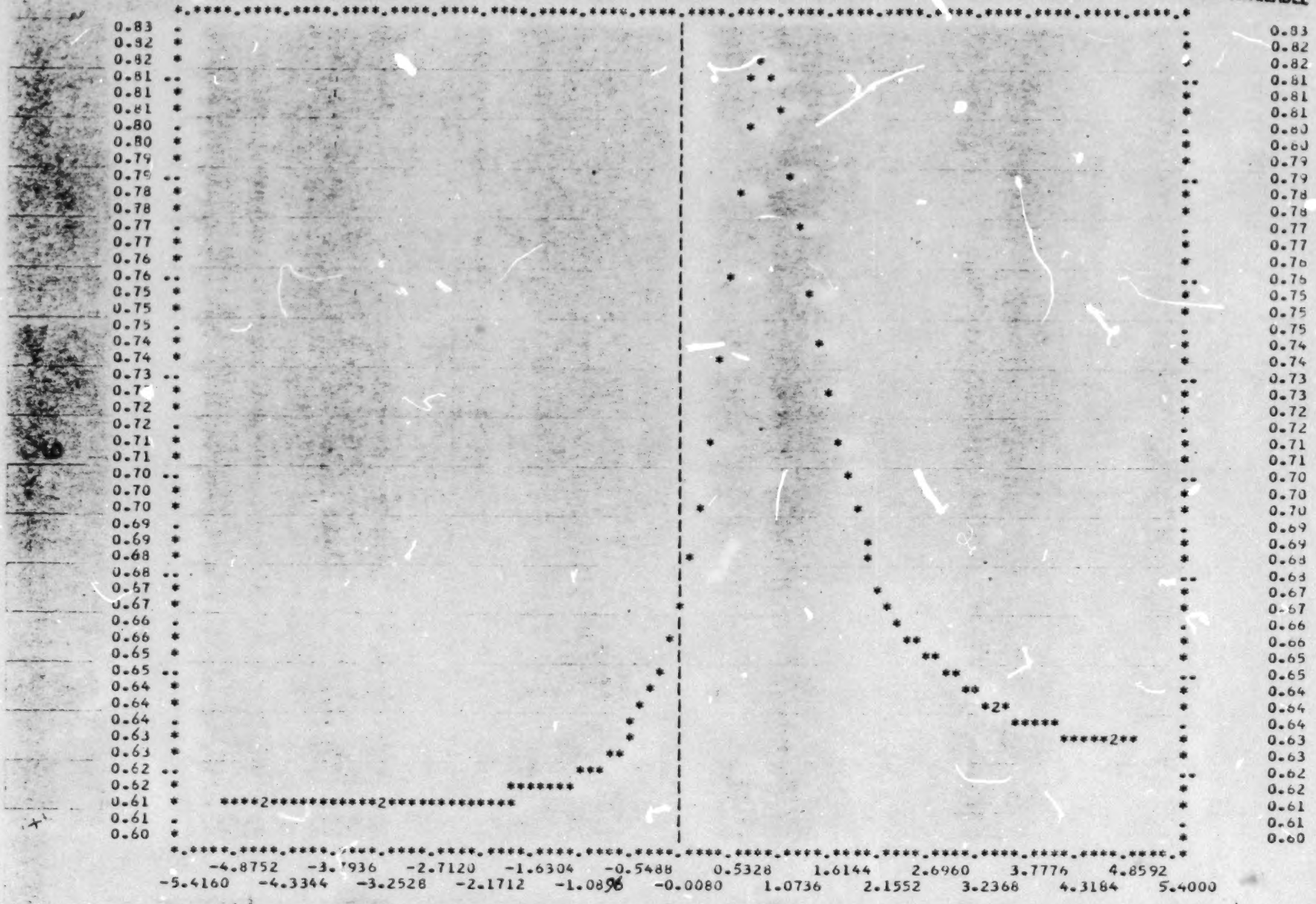
(k)	(R)		(k)	(R)		(k)	(R)	
-0.500000D 01	0.724940D 00	1	-0.160000D 01	0.738140D 00	35	0.180000D 01	0.804600D 00	69
-0.490000D 01	0.724990D 00	2	-0.150000D 01	0.739720D 00	36	0.190000D 01	0.799100D 00	70
-0.480000D 01	0.725050D 00	3	-0.140000D 01	0.741500D 00	37	0.200000D 01	0.794050D 00	71
-0.470000D 01	0.725110D 00	4	-0.130000D 01	0.743520D 00	38	0.210000D 01	0.789440D 00	72
-0.460000D 01	0.725180D 00	5	-0.120000D 01	0.745820D 00	39	0.220000D 01	0.785250D 00	73
-0.450000D 01	0.725260D 00	6	-0.110000D 01	0.748440D 00	40	0.230000D 01	0.781430D 00	74
-0.440000D 01	0.725340D 00	7	-0.100000D 01	0.751420D 00	41	0.240000D 01	0.777960D 00	75
-0.430000D 01	0.725430D 00	8	-0.900000D 00	0.754820D 00	42	0.250000D 01	0.774800D 00	76
-0.420000D 01	0.725530D 00	9	-0.800000D 00	0.758710D 00	43	0.260000D 01	0.771920D 00	77
-0.410000D 01	0.725640D 00	10	-0.700000D 00	0.763150D 00	44	0.270000D 01	0.769300D 00	78
-0.400000D 01	0.725760D 00	11	-0.600000D 00	0.768200D 00	45	0.280000D 01	0.766900D 00	79
-0.390000D 01	0.725890D 00	12	-0.500000D 00	0.773930D 00	46	0.290000D 01	0.764710D 00	80
-0.380000D 01	0.726030D 00	13	-0.400000D 00	0.780420D 00	47	0.300000D 01	0.762700D 00	81
-0.370000D 01	0.726190D 00	14	-0.300000D 00	0.787690D 00	48	0.310000D 01	0.760850D 00	82
-0.360000D 01	0.726360D 00	15	-0.200000D 00	0.795750D 00	49	0.320000D 01	0.759160D 00	83
-0.350000D 01	0.726550D 00	16	-0.100000D 00	0.804550D 00	50	0.330000D 01	0.757590D 00	84
-0.340000D 01	0.726760D 00	17	0.0	0.813970D 00	51	0.340000D 01	0.756150D 00	85
-0.330000D 01	0.726990D 00	18	0.100000D 00	0.823760D 00	52	0.350000D 01	0.754820D 00	86
-0.320000D 01	0.727240D 00	19	0.200000D 00	0.833560D 00	53	0.360000D 01	0.753580D 00	87
-0.310000D 01	0.727520D 00	20	0.300000D 00	0.842680D 00	54	0.370000D 01	0.752430D 00	88
-0.300000D 01	0.727820D 00	21	0.400000D 00	0.851170D 00	55	0.380000D 01	0.751360D 00	89
-0.290000D 01	0.728160D 00	22	0.500000D 00	0.857850D 00	56	0.390000D 01	0.750370D 00	90
-0.280000D 01	0.728530D 00	23	0.600000D 00	0.862440D 00	57	0.400000D 01	0.749440D 00	91
-0.270000D 01	0.728940D 00	24	0.700000D 00	0.864650D 00	58	0.410000D 01	0.748570D 00	92
-0.260000D 01	0.729400D 00	25	0.800000D 00	0.864420D 00	59	0.420000D 01	0.747760D 00	93
-0.250000D 01	0.729900D 00	26	0.900000D 00	0.861950D 00	60	0.430000D 01	0.747000D 00	94
-0.240000D 01	0.730470D 00	27	0.100000D 01	0.857620D 00	61	0.440000D 01	0.746290D 00	95
-0.230000D 01	0.731090D 00	28	0.110000D 01	0.851900D 00	62	0.450000D 01	0.745610D 00	96
-0.220000D 01	0.731790D 00	29	0.120000D 01	0.845270D 00	63	0.460000D 01	0.744980D 00	97
-0.210000D 01	0.732570D 00	30	0.130000D 01	0.838170D 00	64	0.470000D 01	0.744390D 00	98
-0.200000D 01	0.733440D 00	31	0.140000D 01	0.830930D 00	65	0.480000D 01	0.743820D 00	99
-0.190000D 01	0.734420D 00	32	0.150000D 01	0.823820D 00	66	0.490000D 01	0.743290D 00	100
-0.180000D 01	0.735510D 00	33	0.160000D 01	0.817000D 00	67	0.500000D 01	0.742790D 00	101
-0.170000D 01	0.736750D 00	34	0.170000D 01	0.810570D 00	68			

BEST WEIGHTED WRONGS DETERMINATION

Figure 1

DATA SUFFICIENCY PARALLEL FORMS RELIABILITY: WRONGS WT. (X AXIS) VS. RELIABILITY (Y AXIS.

BEST COPY AVAILABLE

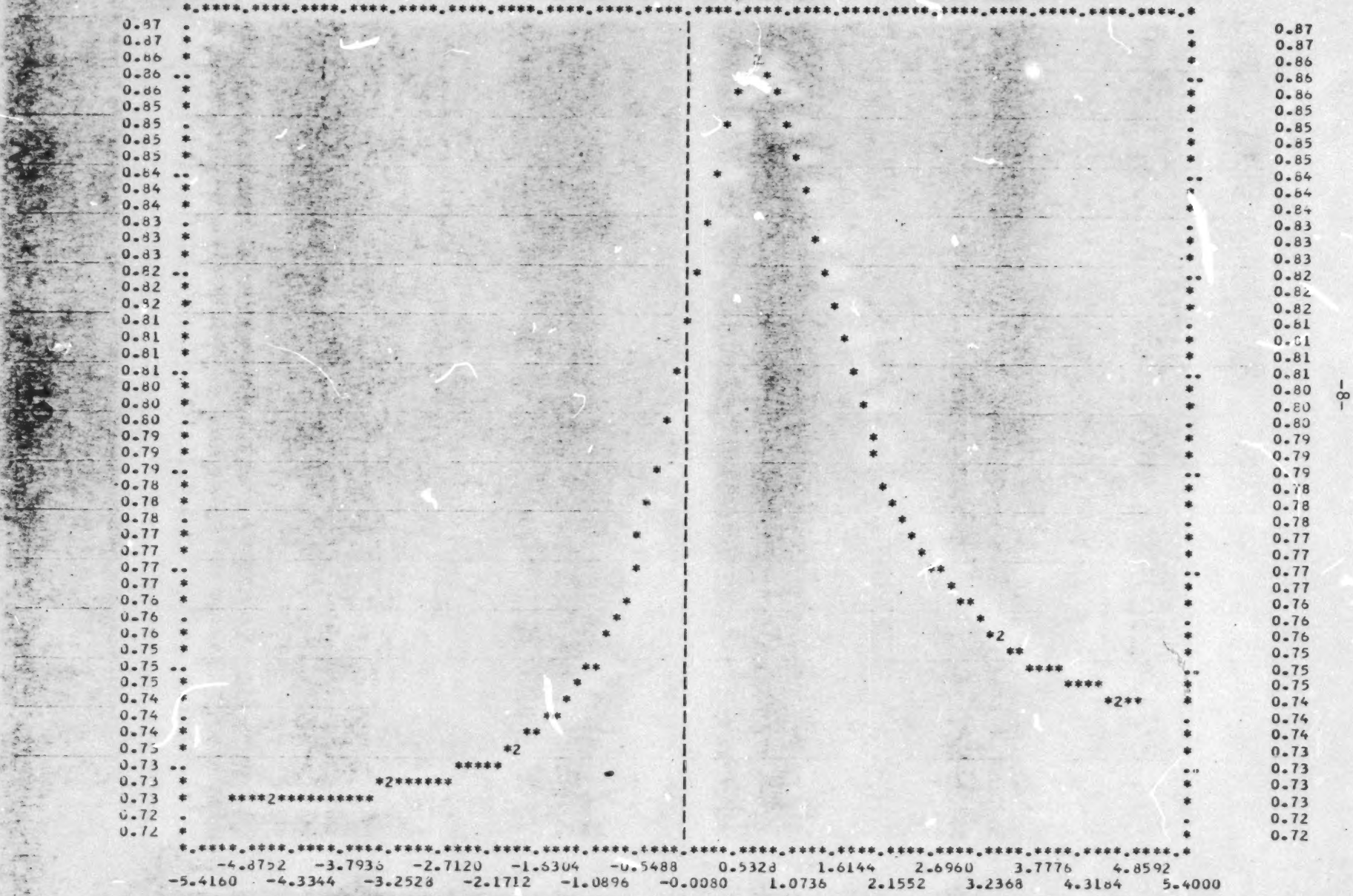


BEST WEIGHTED WRONGS DETERMINATION

Figure 2

MATH PARALLEL FORMS RELIABILITY: WRONGS WT. (X AXIS) VS. RELIABILITY (Y AXIS)

BEST COPY AVAILABLE



to indicate whether or not there is sufficient information in the statements to answer the question. An example would be as follows:

If x is a whole number, is it a two-digit number?

(1) x^2 is a three-digit number.

(2) 10x is a three-digit number.

(A) if statement (1) ALONE is sufficient but statement (2) alone is not sufficient to answer the question asked,

(B) if statement (2) ALONE is sufficient but statement (1) alone is not sufficient to answer the question asked,

(C) if both statements (1) and (2) TOGETHER are sufficient to answer the question asked, but NEITHER statement ALONE is sufficient,

(D) If EACH statement is sufficient by itself to answer the question asked,

(E) if statements (1) and (2) TOGETHER are NOT sufficient to answer the question asked and additional data specific to the problem are needed.

This difference in the item format was accompanied by differences in test content. The data sufficiency material was parallel in content to the College Board SAT, which used about 30% items of this type at that time. The regular math test was parallel to the College Board basic-level achievement test in mathematics. This test has a more advanced content than the Scholastic Aptitude Test.

A third difference between the two tests (in addition to format and content) concerns the development of the half-tests. The data sufficiency test was developed as two separately timed subtests of 25 items each. These were the two half-tests correlated in the current study. The mathematics achievement test was administered with a single time limit and

divided into two half tests consisting of all the odd items and all the even items.

The role of these different factors on the somewhat different outcomes for the two tests is difficult to determine. The maximum value for the data sufficiency test was approximately + 0.90 as a weight for the wrongs. The maximum value for the regular math test was + 0.70 These empirical values contrast with the values of + 0.295 and + 0.585 observed for SAT mathematics subtests in the Fischer and Jackson study.

Table 3 presents the means, standard deviations and intercorrelations for the four half-tests considered in the study. The pattern of intercorrelation is consistent across the two tests. The interhalf reliability of the data sufficiency Rights score was .67 , versus the value of .81 for the math achievement. Similarly the wrongs score for the math achievement test was more reliable, .73 versus .62. The cross-score correlations, $R_1 - W_2$ and $R_2 - W_1$, were - .46 and - .45 for the data sufficiency test and - .34 and - .35 for the math achievement; with similarly lower cross-score correlations for the intratest comparisons ($R_1 - W_1$, $R_2 - W_2$) for the math achievement test.

While omits were not distinguished from Not Reached in the present study, the general trait of omissiveness can be gauged somewhat by considering the numbers of items not responded to in each of the four half-tests studied. The values can be derived from Table 3 as follows:

Table 3

Means, Standard Deviations and Intercorrelations for the Half-Test Scores

Data Sufficiency

	R_1	W_1	R_2	W_2
R_1	1.00	-0.75	0.67	-0.46
W_1	-0.75	1.00	-0.45	0.62
R_2	0.67	-0.45	1.00	-0.79
W_2	-0.46	0.62	-0.79	1.00
Mean	12.15	11.02	12.26	11.24
S.D.	3.82	3.57	3.50	3.36

Math Achievement

	R_1	W_1	R_2	W_2
R_1	1.00	-0.48	0.81	-0.34
W_1	-0.48	1.00	-0.35	0.73
R_2	0.81	-0.35	1.00	-0.51
W_2	-0.34	0.73	-0.51	1.00
Mean	12.51	6.70	12.97	6.60
S.D.	4.43	3.75	4.57	3.67

Average Number of Items Not Responded To

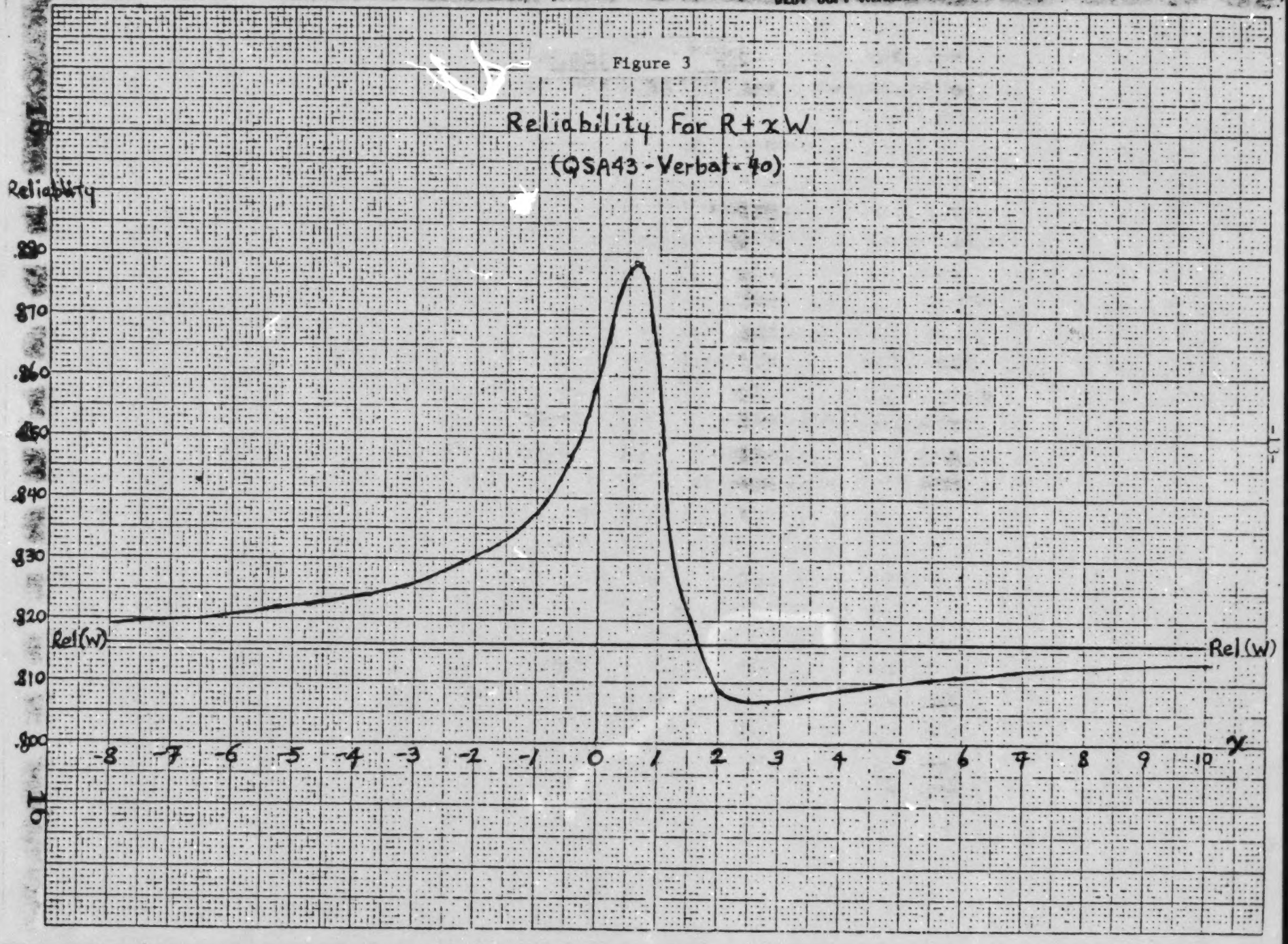
Data Sufficiency Half-Test 1	1.82
Data Sufficiency Half-Test 2	1.50
Math Achievement Half-Test 1	5.79
Math Achievement Half-Test 2	5.42

Clearly, the mathematics achievement test was characterized by a greater tendency to omit. Whether this was due to its greater speededness or to a true lack of knowledge of the material on the part of the subjects cannot be determined from this data. Either is plausible, since it is a characteristic of data sufficiency items that they are processed more rapidly by subjects. Referring to the Fischer and Jackson weights for mathematics tests, which were .295 and .585 , the higher weight was achieved by the section which had a sizable set of data sufficiency items (18 of its 35-item total) and a slightly more generous time allotment, .77 seconds per item versus .72 seconds per item. This suggests that the weight approaches unity as the test is unspeeded. However, the general parity of the number correct on the various half-tests, versus the differences in number wrong, suggests that there may be a greater tendency to give a response to the data sufficiency items, to guess at an answer, than to respond to the mathematics achievement items. This implies a more complex cause for the differences in weight than simply rate of work.

It is interesting to contrast the curves in Figures 1 and 2 with one provided by Fischer and Jackson, presented as Figure 3. In the present study, using empirically determinal curve, there is no suggestion of the minimum point for reliability which is clear in the Fischer and Jackson development. Whether this point would occur outside of the range observed

BEST COPY AVAILABLE

Figure 3



is not clear since no theoretical analysis of the intercorrelations of the tests in this study was undertaken.

Findings of a maximized reliability through a positive weight would seem to indicate that the most reliable aspect of a test performance is the total number of marks which are made. This hardly seems a worthwhile characteristic to focus on, since it would have little implication for validity. However, it is possible that further study of omissiveness would lead to an understanding of the reliability of the two forms of omissiveness: Omits and Not Reached. The best current data on this reliability is available from a study by Flaugher, Melton and Myers (1966), which shows the correlation between a mathematical section of the Scholastic Aptitude Test and each of four other, parallel sections introduced experimentally. The results are summarized in Table 4.

Table 4

Parallel Form Reliability for Four Scores:
Rights, Wrongs, Omits and Not Reached*

	Correlations with Master Form			
Parallel Forms	Rights-Rights	Wrongs-Wrongs	Omits-Omits	Not Reached-Not Reached
1	.790	.700	.628	.452
2	.785	.713	.5[illegible]	.485
3	.776	.720	.648	.464
4	.770	.710	.576	.446

*From Flaugher, Melton and Myers (1966)

This data suggests that the Not Reached score is not as reliable as the Omits score. While this cannot be generalized too broadly, it

bears on the meaning of the positive weight for the maximally reliable composite score. To the extent that the number of omits on parallel forms reflects a reliable tendency not to know a certain proportion of the answers, it is surprising that this would be a more reliable characteristic of an individual than rate-of-work would be. Even with major efforts at content and difficulty parallelism, most parallel forms vary a good deal, so that one would not readily predict that individuals would find comparable numbers of items they would decide not to attempt. Further research seems indicated to clarify the degree to which the Omit response is determined by rate of work.

This paper has confirmed the determination by Fischer and Jackson of a positive weight for the wrongs as a reliability maximizing score. The parallel-forms technique in the present study varied somewhat from the internal-consistency approach which they used. The implications of this weight, as Lord suggests, are that the trait of omissiveness is a reliable one. The source of this reliability and the implications for work on test speededness could be meaningful future areas for research.

Bibliography

Dressel, P.L. Some Remarks on the Kuder-Richardson Reliability Coefficient. Psychometrika, 1940, 5, 305-310.

Ebel, R.L. Essentials of educational measurement. New York: Prentice-Hall, 1972, p 252.

Fischer, F. and Jackson, R. Maximizing Reliability by Utilizing a Proportion of the Wrongs Score. Personal communication from the authors.

Flaugher, R.L., Melton, R.S., Myers, C.T. A Study of the Effects of Item Rearrangement. RB 66-39, Princeton, New Jersey: Educational Testing Service, 1966.

Lord, F.M. Formula Scoring and Number-Right Scoring. Journal of Educational Measurement, 1975, 12, 7-11.

Thorndike, R.L. The problem of guessing. In R.L. Thorndike (Ed.) Educational Measurement. (2nd ed.), Washington, D.C., American Council on Education. 1971, 59-61.